L11: Entry 3 of 7 File: USPT Apr 18, 2000

DOCUMENT-IDENTIFIER: US 6052797 A
TITLE: Remotely mirrored data storage system with a count indicative of data consistency

Abstract Text (1):
Two data storage systems are interconnected by a data link for remote mirroring of data. Each volume of data is configured as local, primary in a remotely mirrored volume pair, or secondary in a remotely mirrored volume pair. Normally, a host computer directly accesses either a local or a primary volume, and data written to a primary volume is automatically sent over the link to a corresponding secondary volume. Each remotely mirrored volume pair can operate in a selected synchronization mode including synchronous, semi-synchronous, adaptive copy--remote write pending, and adaptive copy--disk. Each write request transmitted over the link between the data storage systems includes not only the data for at least one track in the secondary volume to be updated but also the current "invalid track" count for the secondary volume as computed by the data storage system containing the corresponding primary volume. Therefore, once a disaster occurs that destroys the data storage system containing the primary volume, the data storage system containing the secondary volume has an indication of the degree of consistency of the secondary volume. The "invalid tracks" count can be used to determine an appropriate recovery operation for the volume, and can be used to selectively restrict read/write access to the volume when the user decides that synchronization should be required for a write access. Moreover, direct write access to a secondary volume is denied if remote mirroring is not suspended.

Brief Summary Text (14):
This invention features a system which controls storing of primary data received from a primary host computer on a primary data storage system, and additionally controls the copying of the primary data to a secondary data storage system controller which forms part of a secondary data storage system, for providing a back-up copy of the primary data on the secondary data storage system which is located in a remote location from the primary data storage system. For remote copying of data from one storage system to the other without host involvement, the primary and secondary data storage system controllers are coupled via at least one high speed communication link such as a fiber optic link driven by LED's or laser. The high speed communication link also permits one data storage system to read or write data to or from the other data storage system.

Brief Summary Text (19):
At such time, the primary and/or secondary data storage system controller maintaining the list of primary data to be copied updates this list to reflect that the given primary data has been received by and/or copied to the secondary data storage system. The primary or secondary data storage system controllers and/or the primary and secondary data storage devices may also maintain additional lists for use in concluding which individual storage locations, such as tracks on a disk drive, are invalid on any given data storage device, which data storage locations are pending a format operation, which data storage device is ready to receive data, and whether or not any of the primary or secondary data storage devices are disabled for write operations.

Brief Summary Text (21):
In the synchronous mode, data on the primary (R1) and secondary (R2) volumes are always fully synchronized at the completion of an I/O sequence. The data storage

system containing the primary (R1) volume informs the host that an I/O sequence has successfully completed only after the data storage system containing the secondary (R2) volume acknowledges that it has received and checked the data. All accesses (reads and writes) to the remotely mirrored volume to which a write has been performed are suspended until the write to the secondary (R2) volume has been acknowledged.

Brief Summary Text (22):
In the semi-synchronous mode, the remotely mirrored volumes (R1, R2) are always synchronized between the primary (R1) and the secondary (R2) prior to initiating the next write operation to these volumes. The data storage system containing the primary (R1) volume informs the host that an I/O sequence has successfully completed without waiting for the data storage system containing the secondary (R2) volume to acknowledge that it has received and checked the data. Thus, a single secondary (R2) volume may lag its respective primary volume (R1) by only one write. Read access to the volume to which a write has been performed is allowed while the write is in transit to the data storage system containing the secondary (R2) volume.

Brief Summary Text (24):
Another aspect of the present invention provides mechanisms for selectively inhibiting automatic or manual recovery when automatic or manual recovery would be inappropriate. In one embodiment, each write request transmitted over the link between the data storage systems includes not only the data for at least one track in the secondary (R2) volume to be updated but also the current "invalid track" count for the secondary (R2) volume as computed by the data storage system containing the corresponding primary (R1) volume. Therefore, once a disaster occurs that destroys the data storage system containing the primary (R1) volume, the data storage system containing the secondary (R2) volume has an indication of the degree of consistency of the secondary (R2) volume. The "invalid tracks" count can be used to determine an appropriate recovery operation for the volume, and can be used to selectively restrict read/write access to the volume when the user decides that synchronization should be required for a write access.

Drawing Description Text (2):
These and other features and advantages of the present invention will be better understood when read together with the following drawings wherein:

Drawing Description Text (5):
FIG. 3 is a schematic representation of an additional list or index maintained by the system of the present invention to keep track of additional items including an invalid data storage device track, device ready status and write disable device status;

Detailed Description Text (5):
A system in accordance with the present invention is shown generally at 10, FIG. 1, and includes at site A, which is a first geographic location, a host computer system 12 as is well known to those skilled in the art. The host computer system 12 is coupled to a first and primary data storage system 14. The host 12 writes data to and reads data from the primary data storage system 14.

Detailed Description Text (24):
Accordingly, a feature of the data storage system 10 is the ability of a data storage system to control the transfer or copying of data from a primary data storage system to the secondary data storage system, independent of and without intervention from one or more host computers. Most importantly, in order to achieve optimum data mirroring performance, such data mirroring or copying should be performed asynchronously with input/output requests from a host computer. Accordingly, since data will not be immediately synchronized between the primary and secondary data storage systems, data integrity must be maintained by maintaining an index or list of various criteria including a list of data which has not been mirrored or copied, data storage locations for which a reformat operation is pending, a list of invalid data storage device locations or tracks, whether a given device is ready, or whether a device is write-disabled. Information must also be included as to the time of the last operation so that the data may later be synchronized should an error be detected.

Detailed Description Text (29):
In addition to the write pending and format pending bits described above, the data storage system 10 also includes several additional general purpose flags to assist in error recovery. As shown in FIG. 3, invalid track flags 120 including primary bit 122 and secondary bit 124 are utilized and maintained on each data storage device to indicate that the data storage location such as a track, does not contain valid data. Another background task running on the data storage system such as in the service processor or storage system controller constantly checks invalid track bits on each data storage device, and if a bit is found to be set, the copy task is invoked to copy the data from the known good device to the device with the invalid flag track set. Additional flags may be provided such as the device ready flags 126 including bits 128 and 130 which serve to indicate that the device is ready. Similarly, write disable flags 132 may be provided which indicate that a particular primary device or drive 134 or secondary device or drive 136 can presently not be written to. Data can still be copied to the good or enabled drive and then later copied to the disabled drive. If one drive or device is bad, the present invention will set all tracks of that drive as not valid to later cause a copy of all the data.

Detailed Description Text (60):
The semi-synchronous mode is recommended primarily for the long distance option of FIG. 6. The semi-synchronous mode is designed for situations needing high performance at the data storage system containing the primary (R1) volume and tolerating a gap of up to one input/output (worst case) in data synchronization. Although write operations can be held up due to synchronization between primary (R1) and secondary (R2) volumes, read operations continue uninterrupted.

Detailed Description Text (65):
In the first step 401 of FIG. 7, execution branches to step 402 for a read access. In step 402, the channel adapter accesses configuration information, and continues to step 403 if the host is requesting access to a local volume. Preferably, a separate copy of the configuration information is stored in local memory in each of the channel adapters and link adapters. This configuration information identifies whether a volume is local, primary, or secondary, and for each primary or secondary volume, identifies the other volume in the remotely mirrored volume pair.

Detailed Description Text (67):
If the host channel command is requesting data in the primary (R1) volume of a remotely mirrored pair, then execution branches from step 402 to step 405. In step 405, execution branches to step 403 unless the data storage system is in the synchronous mode. For modes other than the synchronous mode, the reading of data from a primary (R1) volume is normally similar to the reading of data from a local volume; in either case, the requested data is fetched without delay from the cache or disk in step 403. Under the abnormal condition of the data being entirely absent from the data storage system due to a disk drive failure, however, a request for data access to a primary (R1) volume can be satisfied by obtaining the requested data from the secondary volume (R2) in the remote data storage system. The handling of such an abnormal condition is discussed below in connection with data recovery procedures.

Detailed Description Text (68):
In step 406, when a remote write is not pending to the secondary (R2) of the requested mirrored volume, execution also branches to step 403 to fetch the requested data from the cache or disk. When a remote write is pending to the secondary (R2) of the requested mirrored volume, however, execution continues to step 407 to suspend the current read task until the remote data storage system acknowledges completion of the pending remote write. Preferably, tasks suspended while waiting for completion of a pending remote write are placed on a first-in first-out (FIFO) queue of suspended tasks, and when the remote data storage system acknowledges completion of the pending remote write, any waiting tasks in queue of suspended tasks are serviced in the order in which the tasks were placed in the queue. Once the remote data storage system acknowledges completion of the pending remote write, and no remote write to the secondary (R2) of the mirrored volume is pending, as tested in step 406, execution branches to step 403 to fetch the

requested data from the cache or disk.

Detailed Description Text (76):
From the control flow in FIGS. 7 and 8, it is clear that when a host writes data to a remotely mirrored volume, the following sequence of events takes place in the synchronous mode: data is written to the cache of the data storage system containing the primary (R1) volume (step 414); an entry is placed in the FIFO link queue for transmission of the data to the data storage system containing the secondary (R2) volume (step 415); the data storage system containing the secondary (R2) volume acknowledges receipt of the data (step 419); the track tables are maintained (step 420); and a device end (DE) signal is presented back to the host that initiated the write request (step 422). In the synchronous mode, all accesses (reads and writes) to the remotely mirrored volume to which a write has been performed are suspended (steps 407 and 412) until the write to the secondary (R2) volume has been acknowledged.

Detailed Description Text (77):
From the control flow in FIGS. 7 and 8, it is clear that when a host writes data to a remotely mirrored volume, the following sequence of events takes place in the semi-synchronous mode: data is written to the cache of the data storage system containing the primary (R1) volume (step 414); an entry is placed in the link FIFO queue for transmission of the data to the data storage system containing the secondary (R2) volume (step 415); a device end (DE) signal is presented back to the host that initiated the write request (step 417); the data storage system containing the secondary (R2) volume acknowledges receipt of the data (step 419); and the track tables are maintained (step 420). In the semi-synchronous mode, read access to the volume to which a write has been performed is allowed (steps 405, 403) while the write is in transit to the data storage system containing the secondary (R2) volume. A second write to the volume is not allowed (steps 411, 412) until the first has been safely committed to the secondary (R2) volume. Thus, a single secondary (R2) volume may lag its respective primary volume (R1) by only one write.

Detailed Description Text (97):
To handle the adaptive modes, a few steps in the flowchart of FIG. 7 are modified. FIG. 9 shows the modifications. In particular, steps 431 to 434 of FIG. 9 are substituted for steps 406 to 407 of FIG. 7, and steps 431 to 434 of FIG. 9 are also substituted for steps 411 to 412 of FIG. 7. It should be apparent that steps 432 and 433 of FIG. 9 are inserted between steps 406 and 407 of FIG. 7 so that when the primary mode is the synchronous mode and a remote write to the volume is pending, the current read task is not suspended in the adaptive mode (step 432) until the number of remote write pending tracks reaches the value of the skew parameter. In a similar fashion, steps 432 and 433 of FIG. 9 are inserted between steps 411 and 412 of FIG. 7 so that when a remote write to the volume is pending, the current write task is not suspended in the adaptive mode (step 432) until the number of remote write pending tracks reaches the value of the skew parameter.

Detailed Description Text (100):
Unless the secondary (R2) volumes are synchronized to the primary (R1) volumes, the data in the secondary volumes may not be consistent. If a local host processor is writing to the primary (R1) volumes at the same time that a remote host processor is reading the corresponding secondary (R2) volumes, the remote processor may read inconsistent data. For example, the local processor may be executing a transaction that transfers $10.00 of a client's funds between two of the client's accounts. The local processor executes a first write that debits the first account by $10.00, and executes a second write that credits the second account by $10.00. If the remote processor reads the secondary volume when only the first write has been written in the secondary volume, and then computes the client's total funds, it will find a loss of $10.00. It is a user responsibility to ensure that the use to which such read-only data is put is consistent with the possibility of data inconsistency. In general, the secondary (R2) volumes should be accessed only after synchronization is achieved by suspending remote mirroring, and waiting until all pending remote writes have been transferred to the secondary volumes.

Detailed Description Text (101):
If a remote host processor should perform a read/write access on an inconsistent

dataset, not only is it possible that the host processor will obtain an inconsistent result, but also the dataset may become further corrupted and made worthless. Unfortunately, in the situation of a disaster that interferes with the data storage system containing the primary (R1) volumes, the best copy of the dataset available may reside in the secondary volumes, and the user may be faced with the difficult decision of whether the dataset should be used for a read/write application, discarded, or in some way repaired with whatever information is available about the past history of the dataset.

Detailed Description Text (104):
The preferred embodiment of the invention addresses these problems in a number of ways. Each write request transmitted over the link between the data storage systems includes not only the data for the track in the secondary (R2) volume to be updated but also the current "invalid track" count for the secondary (R2) volume as computed by the data storage system containing the corresponding primary (R1) volume. Therefore, once a disaster occurs that destroys the data storage system containing the primary volume, the data storage system containing the secondary (R2) volume has an indication of the degree of consistency of the secondary (R2) volume. The "invalid tracks" count can be used to determine an appropriate recovery operation for the volume, and can be used to selectively restrict read/write access to the volume when the user decides that synchronization should be required for a write access.

Detailed Description Text (107):
In the preferred implementation of remote mirroring, primary (R1) and secondary (R2) volumes have particular states that govern host access. A primary (R1) volume is in either a ready state or a not ready state. A secondary (R2) volume is in either a not ready state, a read-only state, or a read-write state. The state of the primary (R1) volume governs access to the primary volume by a host connected to a channel adapter of the data storage system containing the primary volume. The state of the secondary (R2) volume governs access to the secondary volume by a host connected to a channel adapter of the data storage system containing the secondary volume. In other words, the volume state is seen by the host connected to the storage system containing the volume.

Detailed Description Text (110):
(a) Primary Volume Ready

Detailed Description Text (111):
In this state, the primary (R1) volume is online to the host and available for read/write operations. This is the default primary (R1) volume state.

Detailed Description Text (112):
(b) Primary Volume Not Ready

Detailed Description Text (113):
In this state, the primary (R1) volume responds "intervention required/unit not ready" to the host for all read and write operations to that volume. The host will also be unable to read from or write to the secondary (R2) volume associated with that volume.

Detailed Description Text (115):
(a) Not Ready State

Detailed Description Text (116):
In this state, the secondary (R2) volume responds "intervention required/unit not ready" to the host for all read and write operations to that volume. This is the default secondary (R2) volume state.

Detailed Description Text (117):
(b) Read-Only State

Detailed Description Text (118):
In this state, the secondary (R2) volume is available for read-only operations.

Detailed Description Text (119):
(c) Read/Write State

Detailed Description Text (120):
In this state, the secondary (R2) volume is available for read/write operations.

Detailed Description Text (122):
In the event of a disaster that renders all equipment at one site non-operational, secondary (R2) volumes on the mirrored data storage system at the remote site can be made available to a remote host for read-only or read/write operations by issuing commands at the service processor of the data storage system containing the secondary (R2) volumes, or by issuing commands to host remote mirroring software in the remote host. In its default configuration, all secondary (R2) volumes are not ready to the remote host. (These secondary (R2) volumes can also be configured for a read-only state.)

Detailed Description Text (123):
Each secondary (R2) volume has a configurable attribute, "sync required", for selectively preventing a secondary (R2) volume from becoming ready to the remote host if a state change is attempted while it is not synchronized with its primary (R1) volume. If the "sync required" attribute is not enabled, then all specified state changes to the secondary (R2) volume take effect when requested. If the "sync required" attribute is enabled, and if the secondary (R2) volume is not synchronized with the primary (R1) volume and not ready to the remote host at the time of the failure, then the non-synchronized secondary (R2) volume will remain not ready. Regardless of the state of the "sync required" attribute, if the secondary (R2) volume were synchronized with the primary (R1) volume and not ready to the remote host at the time of the failure, then the secondary (R2) volume will assume the specified change of state (read-only or read/write enabled).

Detailed Description Text (124):
Secondary (R2) volumes configured as read-only with the "sync required" attribute enabled can work in their read-only state with the remote host regardless of their synchronization state with the primary (R1) volumes. If an attempt is made to change the state of a secondary (R2) volume to read/write enabled and the secondary (R2) volume is synchronized with the primary (R1) volume at the time of the failure, the state change occurs. If the secondary (R2) volume was not synchronized with the primary (R1) volume, then the state change does not occur and the data storage system reports the non-synchronous state to the remote host.

Detailed Description Text (125):
Turning now to FIG. 10, there is shown a flowchart of the control logic in a channel adapter for restricting the ability of a host to access a secondary (R2) volume in the fashion described immediately above. In a first step 440, execution continues to step 441 if remote mirroring to the secondary (R2) volume has been suspended. When remote mirroring to the secondary (R2) volume has been suspended, writes to the secondary (R2) volume are not accepted from the data storage system containing the corresponding primary (R1) volume. In step 441, execution branches to step 442 if the "sync required" attribute is set for the secondary (R2) volume. In step 442, the requested state change is performed. If the "sync required" attribute is not set for the secondary (R2) volume, then execution continues from step 441 to step 443. In step 443 execution branches to step 442 if the secondary volume (R2) is synchronized with its corresponding primary volume (R1). In other words, execution branches from step 443 to step 442 if the "invalid tracks" count for the secondary volume is zero. If the secondary (R2) volume is not synchronized with its corresponding primary volume (R1), then execution continues from step 443 to step 444. In step 444, execution branches to step 445 if the host is requesting a state change to a read-write state. If so, then in step 445 the state of the secondary (R2) volume is set to "not ready" and the channel adapter reports to the host that the secondary (R2) volume is "not ready." If in step 444 the host was not requesting a state change to read-write, then execution continues from step 444 to step 442 to perform the state change to either "not ready" or read-only, as requested by the host.

Detailed Description Text (126):
If in step 440 remote mirroring was not found to be suspended to the secondary (R2)

volume, then execution branches to step 444 in order to prevent any state change to read-write. However, a state change to read-only or "not ready" is permitted when remote mirroring to the secondary (R2) volume is occurring.

Detailed Description Text (139):
In the automatic mode, if the data is not available in cache during a read operation, then the data storage system reads the data from the primary (R1) volume. If a data check occurs on this device, the data storage system automatically reads the data from the secondary volume. Should one volume in the remote mirrored pair fail, the data storage system automatically uses the other volume without interruption. The data storage system notifies the host with an "Environmental data present" error, and notifies a customer support center of the data storage system manufacturer with an error code designating that the primary or secondary volume has failed. No user intervention is required. When the defective disk device is replaced, the data storage system re-synchronizes the mirrored pair, automatically copying data to the new disk. In a similar fashion, when an outage occurs, e.g., to perform maintenance activity on a remotely mirrored volume for an extended period of time, the primary (R1) volume tracks all updates to its secondary (R2) volume and copies the updated tracks to the other volume when the remotely mirrored pair is re-established. The time it takes to resynchronize the mirrored pair depends on the link path activity, input/output activity to the volume, and the disk capacity.

Detailed Description Text (150):
When the failing or failed disk drive is physically replaced, the data storage system makes the volume(s) on the new disk drive ready, disables the spare, and dynamically copies the ; contents of the other volume in the remotely mirrored pair to the new disk drive. The data storage system returns the spare to its pool, making it available if another remotely mirrored volume (primary (R1) or secondary (R2)) fails in the future.

Detailed Description Text (154):
If all link paths fail between the data storage systems, no data can be written to the secondary (R2) volumes in either data storage system. In an automatic link recovery mode, which is a default configuration, writes from the local host continue to the primary (R1) volumes. All updated tracks are marked so that when the link paths are restored, the data storage system will begin transferring the marked data to the secondary (R2) volumes. In the adaptive copy--write pending mode, all data for the secondary (R2) volume(s) accumulates as invalid tracks in the cache of the data storage system containing the primary (R1) volume(s). In the adaptive copy--disk mode, all data for the secondary (R2) volume(s) accumulates as invalid tracks in disk storage of the data storage system containing the primary (R1) volume(s). In a domino recovery mode, however, the primary volumes become "not ready" to the local host whenever all links fail, in order to maintain synchronization between data storage systems.

Detailed Description Text (157):
The default state for a primary volume is the ready state. If the primary (R1) volume fails, the host will continue to see that volume as "ready", and all reads and/or writes will continue uninterrupted with the secondary (R2) volume in that remotely mirrored pair. However, a domino mode can make the primary volume "not ready."

Detailed Description Text (159):
When enabled for a mirrored volume pair, this mode causes the primary (R1) and secondary (R2) volumes to become not ready to a host if either one of the primary (R1) and secondary (R2) volumes become inaccessible for remote mirroring, for example, due to a disk drive failure or an "all links" failure preventing data transfer between the primary (R1) and secondary volumes (R2). The data storage system responds "intervention required/unit not ready" to a host on all accesses to the "not ready" volume.

Detailed Description Text (160):
To resume remote mirroring after the fault has been corrected, the primary (R1) volume must be made ready again by manual entry of commands to the service processor of the data storage system, or by commands to the host remote mirroring software.

If, however, the primary (R1) or secondary (R2) volume or the links remain down, the primary (R1) volume will immediately become not ready again until the cause of the failure is resolved. If the cause of the failure is resolved and the primary (R1) volume is made ready again, the data storage system containing the primary (R1) volume renotifies its local host that the volume is again ready and brings it online.

Detailed Description Text (163):
When enabled, this mode causes all primary (R1) and secondary (R2) volumes to become not ready if all links fail. When at least one link is reestablished, the primary (R1) volumes must be made ready again by manual entry of commands to the service processor of the data storage system, or by commands to the host remote mirroring software. If, however, the all links remain down, the primary (R1) volumes will immediately become not ready again until a link is established. Once a link is established and the primary (R1) volumes are made ready again, the data storage system containing the primary (R1) volumes renotifies its local host that the primary (R1) volumes are again ready and brings them online.

Detailed Description Text (164):
The all-links domino mode is particularly useful for a cluster of host processors in an open systems environment that uses the link between the processors for sharing data. For example, the shared data would be written by a local host to a primary (R1) volume, transmitted over the link to a secondary (R2) volume, and read by a remote host having read-only access to the secondary (R2) volume. In this situation, it may be desirable to interrupt the application when there is no longer a link. Setting the volumes to a volume domino mode might be too restrictive in this situation, because the shared data could still be written across the link to the secondary (R2) volume even if the corresponding primary volume (R1) would be unavailable.

Detailed Description Text (169):
If step 454 found that there was not a failure to complete a write operation to both the primary (R1) and secondary (R2) volumes, then execution continues to step 458. In step 458, execution branches to step 455 if there was a failure to read a primary (R1) volume. Although a failure to read a primary volume will not in and of itself cause a loss of synchronization between the primary (R1) and secondary (R2) volumes of a remotely mirrored volume pair, such a loss could occur, or become more pronounced, by the time of a following write operation. Therefore, execution branches to step 455 so that if the volume domino mode is not enabled for the primary (R1) volume, then an "intervention required" signal will be presented to the host in step 453 to begin a recovery operation as soon as possible. If, however, the domino mode is not enabled for the primary (R1) volume, and its corresponding secondary (R2) is found to be accessible in step 456, then in step 457 the read operation is completed by reading the secondary (R2) volume.

Detailed Description Text (170):
If step 458 found that there was not a failure to read the primary (R1) volume, then execution continues to step 459. In step 459, execution branches to step 455 if there was a failure to read a secondary (R2) volume. In other words, the secondary (R2) volume was in its read-only state but the read failed, so that the secondary volume would also be unavailable for a write operation during remote mirroring. Again, such a failure to read a the secondary volume will not in and of itself cause a loss of synchronization between the primary (R1) and secondary (R2) volumes of a remotely mirrored volume pair, but such a loss could occur, or become more pronounced, by the time of a following write operation. Therefore, execution branches to step 455 so that if the volume domino mode is not enabled for the primary (R1) volume, then an "intervention required" signal will be presented to the host in step 453 to begin a recovery operation as soon as possible. If, however, the domino mode is not enabled for the secondary (R1) volume, and its corresponding primary (R2) is found to be accessible in step 456, then in step 457 the read operation is completed by reading the primary (R1) volume.

Detailed Description Text (175):
When the data storage system at the local site is ready to be brought back online, recovery can be performed by setting all channel interfaces to online, and

powering-up the local data storage system. The local and remote data storage systems begin synchronizing. When the links synchronize, the primary (R1) volumes begin transferring data to the secondary (R2) volumes. The length of time it takes to resynchronize a full volume depends on the level of activity on the links, the level of activity on the data storage systems, the number of updated tracks (i.e., write pendings or invalid tracks) that need to be copied, link distances between data storage systems, and the size of the volume. The primary (R1) volumes must be in the enabled state for resynchronization to occur. The data storage system sends an operator message to its host when a volume has resynchronized.

Detailed Description Text (179):
In a preferred implementation, as shown in FIG. 12, the application 291 maintains the log file on a remotely mirrored volume pair 291, 293 and the data file 292, 294 on a remotely mirrored volume pair 295, 296 in the data processing system 210. The degree of synchronization between the primary volumes 295 and secondary volumes 296 is selected to guarantee that new data is written to the secondary (R2) log file 293 before the new data is written to the secondary (R2) data file 294. Therefore, the "rolling disaster" scenario is avoided.

Detailed Description Text (180):
The synchronous or semi-synchronous modes, without adaptive copy, will guarantee that data is written to the secondary (R2) copies of the log file 293 and the data file 294 in the same order that the host writes data to the primary (R1) copies 291, 292. Therefore, use of the synchronous or semi-synchronous modes, without adaptive copy, would guarantee that new data is written to the secondary (R2) copy of the log file 293 before the new data is written to the secondary (R2) copy of the data file. However, a less restrictive method is for the application to synchronize the secondary (R2) log file volume 293 just before each transmission of new log file data from the application to the primary data storage system, and to synchronize the secondary (R2) data file volume just before each transmission of the new data file updates from the application to the primary data storage system 214. This less restrictive method ensures that cache overwrite cannot disrupt the sequencing of the log and data file updates in the FIFO link transmission queue.

Detailed Description Text (181):
Turning now to FIGS. 13A and 13B, there is shown an example of a recovery procedure for the system of FIG. 12. If there is a primary system failure such as a complete destruction of the primary data storage system 214, then in the first step 641 of FIG. 13A, the host operating system interrupts the application 292, and the application initiates an application-based recovery program to recover from the secondary (R2) copies of the log file 293 and the data file 294. In step 643, the application inspects time stamps, sequence markers, or beginning/end of file markers in the secondary (R2) copies of the files 293, 294 to determine which one of the two files was last written to. The file last written to can be assumed to be corrupted. If the log file 293 were corrupted, then in step 645 it is discarded and a new secondary (R2) log file is allocated, because the secondary (R2) data file 294 is intact. If the log file 293 were not corrupted, then in step 644 the log file 293 is used to recover the data file 294 by applying to the data file the changes recorded in the log file.

Detailed Description Text (184):
If all links are lost between the primary and secondary data storage systems 214, 246, then processing with the primary (R1) file copies can be suspended until a link is re-established. When the link is re-established, the secondary (R2) file copies can be restored by transferring the pending secondary write data over the link. If the entire data storage system containing the primary (R1) copies is destroyed during the transfer, then it is still possible to recover in the fashion described immediately above for recovering from the destruction of the data processing system having the primary (R1) copies. In other words, the secondary copies of the files are inspected, and the file last written is assumed to be corrupted. If the log file were corrupted, then it can be discarded or re-used, because the data file copy is intact. If the log file were not corrupted, then it can be used to recover the data file by applying to the data file the changes recorded in the log file. This recovery technique still works because in the interrupted transfer of the pending secondary write data over the link, the changes to the secondary (R2) copy of the

data file are always written to the secondary (R2) copy of the log file before the are written to the secondary (R2) copy of the data file.

Detailed Description Text (185):
If all links are lost between the remotely mirrored data storage systems, as tested in step 650 of FIG. 13B, then processing with the primary (R1) file copies can continue in step 651. To avoid the "rolling disaster" scenario, however, the secondary (R2) file copies should not be restored when the link is reestablished in step 652 by transferring secondary write pendings generated since all of the links were lost as in step 654, unless it can be guaranteed, as tested in step 653, that the changes to the secondary (R2) copy of the data file are always written to the secondary (R2) copy of the log file before they are written to the secondary (R2) copy of the data file. If processing with the primary (R1) file copies has continued for any substantial length of time, then it cannot be guaranteed that all updates can be transferred to the secondary (R2) log file before the secondary (R2) data file. Therefore, in this case, execution branches to step 655. In step 655, the secondary (R2) log and data files 293, 294 are saved by configuring them as local copies. Next, in step 656 new, initially empty secondary (R2) files are configured corresponding to the primary (R1) files, and remote mirroring is enabled to copy the primary (R1) log and data files 291, 292 to the new secondary (R2) files. This is an example of a data migration operation upon an active volume, which can be done as described below. Once the new secondary (R2) files have been sufficiently synchronized with the primary files to guarantee that new data is written to the new secondary (R2) log file before the new data is written to the new secondary (R2) data file, recovery has been completed and normal processing may continue. The old, now local secondary file copies can be discarded. However, as tested in step 657, the data storage system containing the primary files could be destroyed during the migration process before recovery has been completed with the new secondary (R2) files. In this case, in step 658, the new secondary (R2) files are discarded and the old, saved secondary (R2) log and data files are restored to their secondary status, and used by the application-based recovery program in steps 643 to 645. This recovery from the old, saved secondary files, however, will recover the state of processing existing just before the all-links failure.

Detailed Description Text (204):
Blocks of cache memory are dynamically allocated when needed to store the logical tracks of data 502. The least-recently-used (LRU) queue 503 contains pointers to cache blocks that are available to be allocated. When a cache block is needed, the pointer at the head of the LRU queue 503 identifies the cache block that should be allocated. If the cache block is needed for a read operation, the pointer is placed at the tail of the LRU queue 503. If the cache block is needed for a write operation, the pointer is taken off the LRU queue 503, and is put back on the LRU queue only when a writeback operation to disk has been completed. The pointer is also kept off the LRU queue 503 for remote write pending in the synchronous, semi-synchronous, and adaptive copy--write pending mode in order to retain the remote write pending data in cache.

Detailed Description Text (205):
The FIFO link transmission queue 504 was described above with reference to step 415 of FIG. 8. In the preferred implementation, this link queue 504 is used in connection with the link buffer 505 in order to prepare information for transmitting commands and data over the link 240 from the link adapter 236 to the remote or secondary data storage system 246 in FIG. 18. The commands transmitted over the link 240 include a write command for a remote write to a secondary (R2) volume in the secondary data storage system 246, and a read command for reading data from a secondary volume (R2) in the secondary data storage system. Each command therefore accesses a single volume. The link queue 504 contains a respective entry for each command that is transmitted over the link 240. Each entry is placed in the link queue 504 by a channel adapter involved in a remote read or write operation, and removed from the link queue by a link adapter that transmits the corresponding command over a remote link to the secondary storage system 246.

Detailed Description Text (213):
Each link adapter scans the link queue 504 in an iterative loop, looking for unlocked entries to service, beginning at the head of the queue. The link adapter

locks the next entry to service, checks the password to determine if the entry is valid, and if so, gets the buffer pointer from the entry, reads the buffer, and builds a job to be executed for transferring data from cache across the link in a direct memory access (DMA) operation. In particular, the link adapter builds a header, and transmits over the link the header, followed by the data, followed by a cyclic redundancy check (CRC). The header, for example, contains a command code such as a code for read or write access, link and command status flags, the logical volume number of the secondary (R2) volume to access, and the invalid track count for the secondary (R2) volume.

Detailed Description Text (219):
In step 544, the link adapter checks whether the entry it is processing is at the head of the link queue, and if not, the link adapter waits until the entry reaches the head of the queue. Then in step 545, the link adapter removes the entry from the head of the link queue, marks the status information of the header with a time stamp or sequence number, and executes the job to send the command over the link, including the header followed by data read from the cache in a direct memory access (DMA) operation, and a cyclic redundancy check. The time stamp or sequence number can be used by the remote data storage system to detect link transmission problems and to write to its cache in proper sequence data from commands received from various links and link adapters despite possible delay of some commands due to link failure. In an alternative arrangement, each link queue entry or corresponding link buffer entry could be marked with a time stamp or sequence number at the time the link queue entry is inserted at the tail of the link queue, so that step 544 could be eliminated. Moreover, in the short distance option configuration having a single link, time stamps or sequence numbers would not be needed, because each command could be transmitted over the link, received, and acknowledged before the next command in the link queue would be transmitted.

Detailed Description Text (236):
Locally mirrored drive (primary (R1) volume) is in a "not ready" state.

Detailed Description Text (238):
Remotely mirrored drive (secondary (R2) volume) is in a "not ready" state.

Detailed Description Text (264):
20--WR Enable (secondary (R2) volume read/write enabled)

Detailed Description Text (265):
10--Not Ready (volume "not ready" to host)

Detailed Description Text (277):
Make a specified primary (R1) volume or range of primary volumes or all primary volumes "ready" to the remote host.

Detailed Description Text (278):
Make a specified primary (R1) volume or range of primary volumes or all primary volumes "not ready" to the remote host.

Detailed Description Text (280):
Enable a specified secondary (R2) volume or range of secondary volumes or all secondary volumes for remote host "read only".

Detailed Description Text (281):
Make a specified secondary (R2) volume or range of secondary volumes or all secondary volumes "not ready" to the remote host.

Detailed Description Text (390):
7. Data storage system message. Valid remote mirroring messages include: DYNAMIC SPARING INVOKED, TARG VOLUME RESYNC W/PRIMARY, PRIMARY VOLUME RESYNC W/SECONDARY, R1 VOL NOT READY STATE, R1 VOL WRITE DISABLED, R2 VOLUME IN NOT RDY STATE, ADAPTER LINK PROBLEM, RESYNC PROCESS HAS BEGUN, ADAPTER LINK OPERATIONAL. Valid migration messages are similar except substitute "DATA MIGRATION COMP ON VOL" for "PRIMARY VOLUME RESYNC W/SECONDARY".

Detailed Description Text (440):
9. Control Unit status. Format is xxx-yy-z. Valid values are: xxx=R/W (read/write mode), xxx=R/O (read only mode), xxx=N/R (not ready mode), xxx=RNR (RDF devices globally not ready), xxx=TNR (secondary (R2) not ready; this status indicates that communication between the remote mirroring pair is currently inactive due to either the link is offline, the link path is physically unavailable or the remote mirroring pair is RDF-Suspended. Use the #SQ LINK command to determine whether the links are online or offline, and the physical connection status of the links), yy=SY (Synchronous mode), yy=SS (Semi-Synchronous mode), yy=AW (Adaptive Copy--Write Pending mode), yy=AD (device is configured for Adaptive Copy--Disk mode), z=I (a secondary (R2) volume to go not ready if the primary (R1) volume (its mirrored device) has invalid tracks on secondary (R2) volume and a state of change has been requested on the secondary (R2) volume), z=D (primary (R1) volume to go not ready if secondary (R2) volume is not ready--Domino mode).

Detailed Description Text (479):
The SC VOL command modifies the status of remote mirroring volumes. This configuration command provides the ability to set the remote mirroring operational mode. All #SC VOL commands require the operator to confirm the action specified, unless this has been disabled by the OPERATOR.sub.-- VERIFY sysparm. This confirmation is necessary as some actions may result in loss of data if performed incorrectly. For example, only one volume in a remotely mirrored pair may be read/write-enabled when the devices are remote mirror suspended. The requirement for confirmation may be bypassed based on the value specified for the OPERATOR.sub.-- VERIFY initialization parameter.

Detailed Description Text (501):
6. Data Migration device status. Valid values are: READY=data storage system device is ready to host; NRDY=data storage system device is not ready to host; NR-MIG=data storage system device is not ready for migration.

Detailed Description Text (533):
1. Set all R2 volumes to a "ready" state to the remote host by typing the following command:

Detailed Description Text (537):
All volumes at the remote data storage system are now available for input/output operations with the host at that site. Before read/write operations can be resumed with the data storage system at the local site, however, all secondary (R2) volumes at the remote data storage system must be set to read-only, not-ready to the host at the remote site, and the resynchronization process established. (Failure to make the secondary (R2) volumes read-only prior to bringing the local data storage system online can result in data corruption and invalid tracks in both the primary (R1) and secondary (R2) volumes.) When the host and data storage system at the local site are ready to be brought back online, perform the following steps:

Detailed Description Text (540):
2. Make all secondary (R2) volumes on the remote data storage system read-only and not ready to the remote host (as per the original configuration) by typing the following commands:

Detailed Description Text (555):
In a normal remote mirroring device relationship, the primary (R1) device may be synchronized with its secondary (R2) device or it may contain updated tracks which the link adapter has not yet sent to the secondary (R2) device (semi-synchronous or adaptive copy state). In addition, in a normal operating environment, the secondary (R2) volume is in a read-only mode. The operator can test recovery procedures by write-enabling the secondary (R2) volumes. To write-enable a secondary (R2) volume, the operator must first suspend remote mirroring operations between the primary (R1) and secondary (R2) volumes, make the devices ready, and then write-enable the secondary (R2) volumes.

Detailed Description Text (559):
(ii) Making Volumes Ready

Detailed Description Text (560):
To make a secondary (R2) volume ready, enter the following command at the host with access to the secondary (R2) volume: #SC VOL,cuu,RDY,dev#. To make all secondary (R2) volumes ready, enter the following command at the host with access to the secondary (R2) volume: #SC VOL,cuu,RDY,ALL.

Detailed Description Text (562):
To write-enable the secondary (R2) volume, enter the following command at the host with access to the secondary (R2) volume: #SC VOL,cuu,R/W,dev#. To write-enable all secondary (R2) volumes, enter the following command at the host with access to the secondary (R2) volume: #SC VOL,cuu,R/W,ALL. Any primary (R1) volume configured with the domino effect option will go RNR (volumes not ready for remote mirroring operation) when remote mirroring operations are suspended. To clear this not ready condition, the operator must disable the domino effect option on those "not ready" volumes, and then enable those devices for remote mirroring operation using the RDF-RDY action with the #SC VOL command.

Detailed Description Text (567):
1. Make the secondary (R2) volume(s) on the data storage system read-only by typing the following command at the host with access to the secondary (R2) volume(s): #SC VOL,cuu,R/O[,dev#.linevert split.,ALL].

Detailed Description Text (568):
2. Make the secondary (R2) volume(s) on the data storage system not ready by typing the following command at the host with access to the secondary (R2) volume(s): #SC VOL,cuu,NRDY[,dev#.linevert split.,ALL].

Detailed Description Text (573):
2. Make the secondary (R2) volume(s) on the data storage system read-only by typing the following command at the host with access to the secondary (R2) volume(s):

Detailed Description Text (575):
3. Make the secondary (R2) volume(s) on the data storage system not ready by typing the following command at the host with access to the secondary (R2) volume(s):

Detailed Description Text (591):
2. Make the secondary (R2) volume(s) on the data storage system read-only by typing the following command at the host with access to the secondary (R2) volume(s):

Detailed Description Paragraph Table (1):

Possible Actions: (R1 = primary volume, R2 = secondary volume)	Action	Valid Volume Type	Description
	R/W	R2	Make secondary (R2) device(s) read and write enabled. This allows a secondary (R2) to be written to from the channel. Please note that if you write to the secondary (R2) device, you should perform testing and recovery procedures.
	R/O	R2	Make secondary (R2) device(s) read-only. When a secondary (R2) volume is in this status, any attempt to issue a write from the channel produces an input/output error.
	RDY	R2	Make secondary (R2) device(s) ready to the host.
	NRDY	R2	Make secondary device(s) not ready. In this state, the secondary (R2) volume responds "intervention required" to the host for all read and write operations to that volume. This is the default state for a secondary (R2) volume.
	SYNC	R1	Set primary (R1) device to the synchronous mode. This is a remote mirroring mode of operation that ensures 100% synchronized mirroring between the two data storage systems.
	SEMI-SYNC	R1	Set primary (R1) device to the semi-synchronous mode. This is an remote mirroring mode of operation that provides an asynchronous mode of operation.
	DOMINO	R1	Enable volume domino mode for primary (R1) device. This ensures that the data on the primary (R1) and secondary (R2) volumes are fully synchronized at all times in the event of a failure.
	NDOMINO	R1	Disable volume domino mode for primary (R1) device. During this default operating condition, a primary (R1) volume continues processing input/outputs with its host even when an remote mirroring volume or link failure occurs. These failures cause loss of primary (R1) and secondary (R2) synchronization. When the failure is corrected, the devices begin synchronizing.
	RDF-RDY	R2/R1	Set volume ready to the host for remote mirroring operation. This

action is valid for both primary (R1) and secondary (R2) volumes. RDF-NRDY R2/R1 Set volume not ready to the host for remote mirroring operation. This action is valid for both primary (R1) and secondary (R2) volumes. ADCOPY-WP R1 Enable adaptive copy - write pending function for primary (R1) device. When this attribute is enabled, data storage system acknowledges all writes to primary (R1) volumes as if they were local volumes. NADCOPY R1 Disable Adaptive Copy Function for primary (R1) device. Please note that when switching from adaptive copy - disk mode to adaptive copy - write mode or from adaptive copy - write mode to adaptive copy - disk mode, this command must first be used before setting the new adaptive copy mode. Please note that when this command is issued to remove a device from adaptive copy mode, the state change will not take place until the volumes are synchronized. ADCOPY-DISK R1 Place the specified device(s) in adaptive copy disk mode. ADC-MAX R1 Set the adaptive copy maximum skew value for the device(s). Example: #SC VOL,F00,ADC-MAX,,80. The maximum skew value may be specified in the range of 1-999999. This command may only be entered when the device is in one of the supported adaptive copy modes. Setting the skew value too high in Adaptive Copy - Write Pending mode could result in excessive cache use adversely affecting data storage system performance. RDF-SUSP R1 Suspend remote mirroring operation on specified device. If the device is already suspended, this action is ignored. RDF-RSUM R1 Resume remote mirroring operation on specified device. This action is only valid if the device was previously suspended via a successful RDF-SUSP action or INVALIDATE action. VALIDATE R1/R2 Make all tracks for a primary (R1) volume valid on a secondary (R2) volume. When SYNCH.sub.-- DIRECTION=R1>R2 this action code makes all tracks from a primary (R1) volume valid on secondary (R2) volumes. When SYNCH.sub.-- DIRECTION=R1<R2 this action code makes a primary (R1) volume not ready and prepares it to be re-synched from the secondary (R2) volume using RDF-RSUM. It makes all tracks for a secondary (R2) volume valid on the primary (R1) volume. INVALIDATE R1 Make all tracks invalid for a secondary (R2) volume on a primary (R1) volume. When resynchronization begins, all primary (R1) volume tracks are copied to the secondary (R2) volume.

US Reference Patent Number (87):
5544347

Other Reference Publication (42):
Storagetek 2Q Earnings Down, Iceberg Ready for Testing by Jim Mallory, Newsbytes, Jul. 15, 1993.

CLAIMS:

11. A program storage device readable by a data storage system, said program storage device encoding a program for execution by the data storage system for controlling transmission of remote copy data over a data link from the data storage system to remote data storage, wherein the program is executable by the data storage system for maintaining a count of a number of data storage locations which are invalid on the remote data storage, and for transmitting over the data link to the remote data storage the count of the number of data storage locations which are invalid on the remote data storage.

14. A program storage device readable by a data storage system, said program storage device encoding a program for execution by the data storage system for controlling use of remote copy data received by the data storage system from a data link, wherein the program is executable by the data storage system to store the remote copy data in data storage, to receive from the data link a count of a number of data storage locations which are invalid on the data storage, and to access the count to determine an appropriate recovery operation after a failure.